Using Narrative to Enable Longitudinal Human-Robot Interactions*

Michael A. Goodrich
Computer Science Department
Brigham Young University
Provo, UT

Jacob W. Crandall
Computer Science Department
Brigham Young University
Provo, UT

Mayada Oudah
Social Science Department
New York University Abu Dhabi
Abu Dhabi, UAE

Najma Mathema
Computer Science Department
Brigham Young University
Provo, UT

ABSTRACT

New designs and metaphors for human-robot interaction need to be created to support interactions between humans and robots that occur over many days, weeks, months, or years. This position paper introduces a dynamic trajectory metaphor for long-term interaction as a path through memory, mental states, and dispositions. Using this metaphor, the paper outlines how the deliberate construction of narrative can be used to shape the trajectory, avoiding undesirable equilibria such as the overuse, disuse, or misuse of robots caused by miscalibrated trust. The metaphor is illustrated to long-term service robots, in which humans have a longitudinal and local relationship with "their" service robot as well as a longitudinal and distant relationship with a robot service provider. Creating deliberative and adaptive narratives in such context will likely require learning from vast distributed data models, and will require both general learning components that reside at some service bases (which will rely on data collected from multiple robots), and specialized learning components localized within a particular set of locations and group of humans. We review some past work that illustrates some basic components for the artificial intelligence required to create such narratives, using so-called "cheap-talk" in repeated games as an example of how communication can help abstract from episodic interactions to narrative-based modes of thought.

1 INTRODUCTION

Long-term human-robot interaction (HRI) should span months or years. Such long-term interaction will necessarily shift interaction from traditional fixed roles like supervisor and mentor [9, 11] to more relationship-based roles like "durative assistant" or collaborator [8]. This position paper uses the notion of narrative to create a framework for understanding sustainable, meaningful, and productive long-term HRI. The narrative framework is applied to a specific HRI instance: longitudinal HRI between a human and a robot, where the robot is supported by big data learning and analytics from a robot service provider.

New metaphors and artificial intelligence (AI) algorithms for HRI are required to enable long-term interaction. This position paper introduces two new metaphors: (a) long-term interaction as a dynamic trajectory through memory, mental states, and dispositions, and (b) narrative as a means for shaping this dynamic trajectory. The paper also proposes a model for the "characters" of the narrative, using a distributed robot service model. Finally, the paper reviews how "cheap talk" in repeated games illustrates some of the first steps for what the AI must look like to enable longitudinal, narrative-based interactions.

Long-term interaction necessitates a pivot from episodic measures of interaction quality to longer-term measures. The shift from episodic measures to longer-term measures means that we are not trying to maximize engagement, minimize workload or maximize transparency in any traditional sense; these local optimizations are episodic means to achieve long-term objectives. For example, maximizing engagement can be done through gamification – setting up artificial reward structures using things like giving points, enabling upgrades in abilities or access to information, or unlocking new characters or worlds.

Measures of the quality of longitudinal interactions tend toward more comprehensive metric types. For example, "[With] enhanced autonomy. ... [i]ndividuals are using their newly expanded practical freedom to act and cooperate with others in ways that improve the practiced experience of democracy, justice and development, a critical culture, and community." [2, Chap.1]. Similar standards for successful longitudinal interaction complement trends in defining success in user-system design, "users' efficiency, safety, and satisfaction have expanded to also include issues like meaning, engagement, and fulfillment"[6, p.16]. Measurement classes like meaning, justification, fulfillment, and community help highlight how long-term interaction is fundamentally different from short-term interaction. We now discuss the guiding metaphor for this paper.

2 METAPHOR: LONG-TERM INTERACTION AS A DYNAMIC TRAJECTORY

Long-term HRI requires new designs and metaphors to support interactions between humans and robots that occur over many days, weeks, months, or years. We propose the following metaphor for such long-term interaction. Long-term interaction is a dynamic trajectory through short-term episodic memory, long-term memory (including conceptual, declarative, procedural, and autobiographic), mental states (including beliefs, desires, and intentions), and subjective dispositions (including trust, frustration, satisfaction, and sentiment). We recognize that this statement is vague, and a challenge for future work is to make this metaphor more precise and

Proceedings of HRI2018 Workshop on Longitudinal Human-Robot Teaming, Chicago, March 2018.

practical. Another challenge is that the trajectory should probably be thought of as passing through an even larger space or "landscape" as follows: the human experience in a long-term interaction with a robot can be thought of as a trajectory through cognitive, inter-subjective, emotional, social, cultural, physical, economic, sociological, and organizational space.

Dynamic trajectories can reach both desirable and undesirable equilibria. We choose the concept of *trust* in HRI to illustrate this trajectory. Lee and See's excellent overview of trust [14] builds on Parasuraman and Riley's identification of three undesirable steady states associated with trust: misuse, disuse, and abuse [20]. Lee and See emphasize misuse and disuse. Misuse includes over-reliance on an automated system, accepting decisions and actions from the system without appropriate evaluation, and a human does not feel capable of determining when to override or intervene. Disuse includes under-reliance or abandonment of the system, often because of errors committed by the system. There is much contemporary work on "calibrating" trust, which often means cultivating expectations and interaction frequencies that match performance and allow automation errors to be caught and fixed [10].

A "trust trap" is an undesirable equilibria that can appear in the dynamic trajectory as the service robots learn. They are undesirable because they are inefficient in the Pareto optimal sense, meaning that there exist other equilibria that are better across most or all relevant performance objectives that humans might value. Specifically, trust traps will lead to misuse or disuse of the robots. Conflicts and failures will arise between the people the robots serve and the robots as robots make mistakes, etc. Since such conflicts affect trust [10], inevitable conflict or failure makes a trust trap a real possibility in long-term interaction.

3 DISTRIBUTED ROLES AND DATA

Embodied agents like Siri, Alexa, Cortana, etc. address this problem using big-data approaches in which data collected from thousands of similar devices used by thousands of users is used to update and improve the capability of all such embodied agents. At the same time, these agents can adapt to the particulars, peculiarities, and preferences of an individual human. Thus, there exist two parallel adaptations for these embodied agents: a generalized learning component driven by big-data that occurs away from the device and a specialized learning component that adapts to a particular human or small group of humans.

The specialized individual adaptations and the generalized data-driven adaptations lead to different ways of providing benefits to humans. We adapt the following quote, written in the context of using networked information to enable human values, to longitudinal HRI (replacing "networked information economy" with "longitudinal HRI" in the quote) [2, Chap.1]

> [Ideally, longitudinal HRI] improves the practical capacity of individuals along three dimensions: (1) it improves their capacity to do more for and by themselves; (2) it enhances their capacity to do more in loose commonality with others, without being constrained to organize their relationship through a price system or in traditional hierarchical models of social and economic organization; and (3) it improves the capacity of individuals to do more in formal organizations that operate outside the market sphere.

Of course, items (2) and (3) above are only implicit in the model proposed in this paper – the "loose commonality" and "organizations ... outside the market sphere" are enabled through data-driven learning managed by the service provider.

For longitudinal HRI, we consider long-term interactions between a robot (who is provided by a robot service) and a person (who receives services from the robot). Because the robot receives updates from the robot service and because humans will be aware of this service provider, it is likely that humans will ascribe intention not only to the robot but also to the service or designers of the robot [6]. Hancock et al. say it this way: "Although we are often frustrated with technological shortcomings and failures and express our frustration accordingly, at heart, we know we are dealing with the residual effects of a remote human designer" [10, pg. 523]. The way Crilly [6] illustrates the mental models associated with human, robot, and designer is shown in Figure 1.

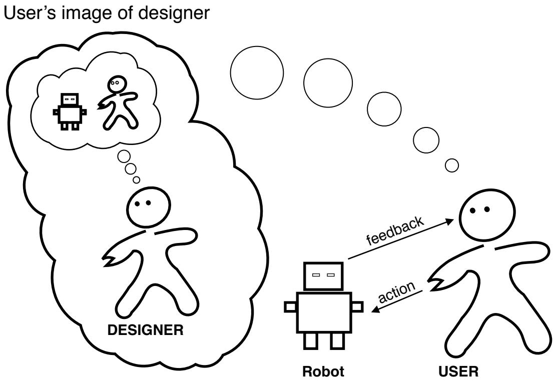

Figure 1: Crilly's mental models associated with human, robot, and designer. Adapted from [6].

For example, the robot service may provide service robots to households. These service robots may provide any number of services over a long period of time. As such, in each household, the robot and robot services must establish and maintain a successful long-term relationship with members of the household, with the relationship modulated by attributed intention and capabilities to the robot service.

Figure 2 illustrates the model for learning and adaptation that we propose for long-term HRI: a generalized learning component that resides at some service base and relies on data collected from multiple robots and multiple humans, and a specialized learning component localized within a particular set of locations and particular group of humans.

4 NARRATIVE

We propose that the robot's behavior and speech acts, as well as the service provider's interactions with the customer should be embedded into a coherent narrative. Part of the narrative is to

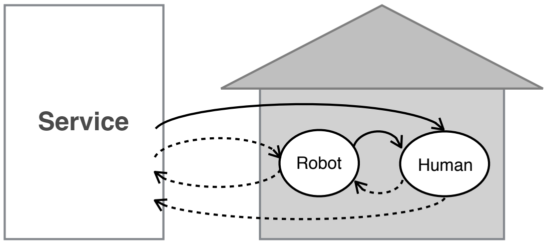

Figure 2: A robot service provides robots to households. The image here depicts the lines of communication (shown with arrows) between the service and a single household. Solid arrows indicate communication channels that the service can use to create a narrative that helps establish and maintain a successful relationship with the human customer.

organize and summarize previous interactions in a way that allows the human to perceive and comprehend robot/service provider's intentions. Another part of the narrative is to embed future possible interactions into a narrative that allows the human to project and believe in the utility of future interactions, hopefully in such a way that the overuse, misuse, and disuse trap is avoided.

We use Polletta et al.'s [21] description of how Labov and Waletsky [13] define narrative: "A narrative is an account of a sequence of events in the order in which they occurred to make a point". Not all narrative structures reveal plot in a linear sequence, but since human experience proceeds in time, order is an acceptable element. The generalized-learning/specialized-adaptation model provides an approach that can enable a robot to learn enough to make it possible for long-term HRI to be useful and extensible, but it does not solve the problem of avoiding the overuse, misuse, or disuse trap. A complementary model is required.

4.1 Why Narrative?

Narrative is compatible with long-term interaction because it enables a natural and coherent representation of the world. "We think in story. It's hardwired in our brain. It's how we make *strategic sense* of the otherwise overwhelming world around us. Simply put, the brain constantly seeks meaning from all the input thrown at it, yanks out what's important for our survival ... and tells us a story about it, based on what it knows of our past experience with it, how we feel about it, and how it might affect us " [7, Chapter 1, emphasis added]. The story provides meaning and context for the shared trajectory of human and robot.

Szilas describes the narrative hypothesis as meaning that narrative is a basic mental construct by which a "large class of real-world happenings [are interpreted]" [22, pg.134]. Regardless of whether this hypothesis is correct, we propose that narrative can shape how interaction episodes, with key features represented in the human's episodic memory, can be organized into a variant of autobiographical memory. In this variant, we trade the autobiographical memory of the human for an intentional relationship memory, conceptualized and organized using narrative cues and context provided by the robot.

Deliberately using narrative to organize episodes therefore hypothetically helps long-term interaction to evolve in such a way that undesirable equilibria are avoided. Narrative can give context for conflict and failure, as well as a path to resolving conflict and failures. The key task of the proposed research is to construct learning algorithms and design interaction episodes to provide narrative-based explanation. This requires a construction of ML algorithms that take feedback experienced in episodic interactions, perhaps over many interactions with many robots and many humans, and constructs a plausible narrative that carries the long-term relationship forward through desirable equilibria.

Stated another way, the task is to design an intentional narrative that avoids or recovers from undesirable equilibria, such as the misuse, disuse, and overuse trap. Events will arise when the robot will do something unexpected or that fails to meet expectations. Has the narrative prior to those events produced expectations about these failures and, shaped by intent and an understanding of the intent holders, opened a path for rebuilding trust and overcoming conflict? "Conflict is story's lifeblood ... but conflict [must be] *specific to the protagonist's quest*" [7, Ch.1, emphasis in original]. The protagonist can be either the robot, the human, or the service provider/designer, and the quest is an evolving element of the narrative to support quality of life for the human.

4.2 Bruner's Landscapes

Bruner identifies two parts to narrative: "One is the landscape of action, where the constituents are the arguments of action: agent, intention or goal, situation, instrument, something corresponding to a "story grammar." The other landscape is the landscape of consciousness: what those involved in the action know, think, or feel, or do not know, think, or feel" [3, Chapter 2]. We begin with the action landscape by localizing specific robot-human interactions or exchanges within the context of an episodic memory organized according to a plot. We will then describe elements of the consciousness in terms of intention and attribution.

Bruner's landscapes provide a framework for designing and evaluating narratives for real real systems. Stated another way, these landscapes provide specific areas of future work and can be used to extend computational narrative architectures [15].

4.2.1 Action Landscape: Episodes and Plot. The first element of Bruner's *action landscape* that we consider is an interaction episode. Story-telling traditionally breaks a narrative that evolves over time or even multiple generations into chapters. Each chapter captures key ideas from an interaction episode (including conflict or changes in intention or information). Each chapter is an episode that gives context and meaning in how the plot-driven narrative evolves [7].

The relationship between a narrative and interaction episodes is complex for a number of reasons. Most obviously, a robot designer may refer to an hour-long interaction between a human and a robot as an *interaction episode*, but this is much different from the notion of *episodic memory* as used in cognitive science. This section reviews how interaction episodes can be related to episodic memory and then aggregated into what we call a *relationship biography*, beginning with a description of episodic memory.

In his great primer on working memory, Baddeley describes episodic memory as follows, "[Episodic memory is a buffer that] is capable of holding multidimensional episodes or chunks, which may combine visual and auditory information possibly also with smell and taste. It is a buffer in that it provides a temporary store in which the various components of working memory, each based on a different coding system, can interact through participation in a multidimensional code, and can interface with information from perception and long-term memory" [1]. Episodic memory in this sense is an integration locus, where the visual short-term memory store, the phonological loop, and long-term memory blend to combine bottom-up perceptual processing with the top-down semantic meaning.

Conway proposes a hierarchical model that uses episodic memories as the basic building block [4]. Two key ideas are used in this model. First, even though episodic memories are short-term, meaning that the ability for humans to recall intentionally episodic memories decreases over time, contextual cues can trigger the recollection of these episodic memories over a period of days and sometimes longer. Thus, episodic memories are short-term in terms of intentional recall but longer-term in terms of their responsiveness to cued recall. Second, episodic memories are augmented with conceptual knowledge to form more long-lasting memories. "There is ... evidence that episodic memories are crucial in the acquisition of new knowledge and learners may pass through a phase during which knowledge is gradually abstracted from episodic memories in the process of becoming part of more general long-term conceptual knowledge" [4, pg.2307]. Episodic memories, linked with conceptual knowledge, can be abstracted and linked to form a frame, and frames and events can be linked with notions of a "conceptual self" to form autobiographical knowledge. Thus, Conway's model suggests that episodic memories, suitably associated with conceptual knowledge, are the building block from which autobiographical knowledge is constructed. Some computational models use abstraction and conceptual linking to promote efficient memory retrieval, though such work does not go all the way toward forming autobiographical memory [12]. Another cognitive architecture seeks to connect "episodic memory and procedures [akin to Conway's conceptual knowledge] can be redefined in terms of narrative structures" [15]. The idea of this architecture is that narrative is a fundamental way of "storing material in memory".

In terms of the narrative trajectory, episodes are temporally limited interactions between a human and a robot, localized in time and space and in a specific context. We propose that the human way of chunking experiences and the evolving robot capability of engaging in experiences can best be coordinated by deliberating choosing local human-robot interactions so that they move a plot (see the next section) along. The overall narrative is punctuated by localized interaction episodes that shape shared cognition, shared experiences, empathy, and the space of possible shared and individual intentions. The localized interaction may be successful in terms of task achievement or it may fail, but it will affect the trust dynamic (e.g., trust traps) and expectations about future interactions.

The second element of Bruner's action landscape that we consider is plot. Like a reader of a story, a human engaged in long-term interaction with a robot may ask, "1. Whose story is it? 2. What's happening here? 3. What's at stake?" [7, Chapter 1]. Thus, the narrative provided to the human by the robot and robot service must touch on plot. Bruner states that "[t]he plot is how and in what order the reader becomes aware of what happened. ...[T]he 'same' story can be told in different sequence[s]" [3, Chapter 1]. *Polleta et al.* further states that "only relevant events are included in the story, and later events are assumed to explain earlier ones. The causal links between events, however, are based not on formal logic or probability but on plot. Plot is the structure of the story"[21, p.111].

A challenge in communicating plot in longitudinal HRI is that, unlike story-telling with a known plot, longitudinal HRI requires the plot to evolve as events unfold and as new knowledge, skills, and awareness are acquired. As such, we hypothesize that the effective communication of plot in longitudinal HRI will have two components. First, the narrative should comment on and interpret events as they unfold. Second, the narrative should foreshadow how future interaction episodes will unfold. Importantly, this narrative should help the human identify how the behavior of the human, the robot, and the robot service can and will influence the future states and events of the interaction.

4.2.2 Consciousness Landscape: Intention, Attribution, and Repair. In addition to the action landscape, Bruner also identifies a consciousness landscape. The first element of Bruner's consciousness landscape is intention, and in particular the intention of the characters or subjects of the narrative. For example, Bruner writes that "narrative deals with the vicissitudes of human intentions" [3, Chapter 2], making these vicissitudes the common theme that Bruner says is shared by (almost all) meaningful stories. Malle states that the "concept of intentionality is essential to people's descriptions and explanations of behavior" [17, pg. 116]. In longitudinal HRI enabled by a robot service, intention can be attributed by the human(s) to both the robot and the robot service provider or designer [6].

This line of thought re-focuses the design of human-robot systems away from the problem of creating perfect robot autonomy to the problem of identifying failures, attributing these failures to proper sources, and then creating a believable narrative that repairs and re-calibrates human trust in the system, while communicating a believable plan of future improvements. Much work needs to be done in this line of reasoning, but learning and shaping intention (the robot's, the robot service provider's, and the human's) is one way to abstract episodes into plots.

The second and third elements of Bruner's consciousness landscape are attribution and repair, respectively. Work on attribution has studied asymmetries in how people explain their own behavior as opposed to how they explain the behavior of others. Jones and Nisbett initially presented evidence that people are more likely to cite situation causes to explain their own behavior, but are more likely to attribute the behavior of others to personal disposition. Malle [16] later demonstrated through a meta-analysis that this theory did not fully hold, but verified that asymmetries in how people attribute blame to themselves and others do exist [18].

Knowledge of the asymmetries in attribution of blame highlight a number of questions relevant to how a service robot and robot service should address the attribution of blame when the robot makes mistakes or otherwise fails to deliver what the human expects. These questions include: What attributions of blame should the robot and robot service voice in their narratives? How should these attributions of blame be communicated? And finally, who has the responsibility for overcoming the problem?

	C	D
C	60, 60	0, 100
D	100, 0	20, 20

Table 1: A payoff matrix defining the well-known Prisoner's Dilemma. In each round, Player 1 selects the row, while Player 2 selects the column. The resulting cell of the matrix specifies the payoffs obtained by players 1 and 2, respectively, in the round.

This latter question moves us to the next item in Bruner's landscape of consciousness: repair. Errors or failures by the robot to satisfy the human's intent are likely to lead distrust, which in turn can lead to misuse or disuse of the robot in the future, and attributions of blame can potentially even exacerbate the problem. In such situations, it is important for the robot or robot-service to create a narrative of how these errors and failures can be overcome in the future. Such repairs to the relationship involve both communicating a plan for how such failures will be overcome, and subsequently following through with the plan in a satisfactory nature.

Much work needs to be done in this area to create an operational narrative-based interaction method, but Bruner's landscapes provide a framework under which designs can be created and evaluated.

5 BUILDING NARRATIVE AI TO ENABLE LONG-TERM HRI

We hypothesize that narrative can play a critical role in enabling long-term HRI. To support this claim, we appeal to past work studying human-human and human-robot interactions in repeated games. This past work highlights how people and robots can enhance and preserve relationships using cheap talk. We then relate how this cheap talk expresses important, albeit primitive, parts of narrative by relating it to the elements of narrative articulated by Bruner.

5.1 Modeling Long-Term Relationships as Repeated Games

A long-term relationship between a human and a robot is created by a sequence of interactions between the human and robot across which both entities seek to achieve their objectives. These interactions can somewhat abstractly be modeled using repeated games [5]. In a repeated game, players engage in a series of rounds (or episodes) of interaction. In each round, each player independently selects an action. The resulting joint action produces an outcome described by a payoff to each player for that round. Over the course of the repeated game (i.e., throughout all rounds of the game), each player seeks to maximize its payoffs.

As an example, consider the well-known prisoner's dilemma. In a repeated prisoner's dilemma game, the same two players repeatedly play the bi-matrix game depicted in Table 1. In each round, the players select a row or column, respectively, of the payoff matrix, which produces the round's payoff for both players. Thus, the dynamics of the actions chosen by the players over time determines in large part the quality of the player's relationship. While each player individually benefits in each round by *defecting* (action D), this results in a low payoff to both players. Thus, if the players repeatedly defect against each other, they both will likely want to discontinue the relationship since it is not profitable. However, if the players could somehow convince each other to cooperate (action C) in each round, they would both receive much more benefit from the relationship, and would be more likely to want to continue to interact with each other.

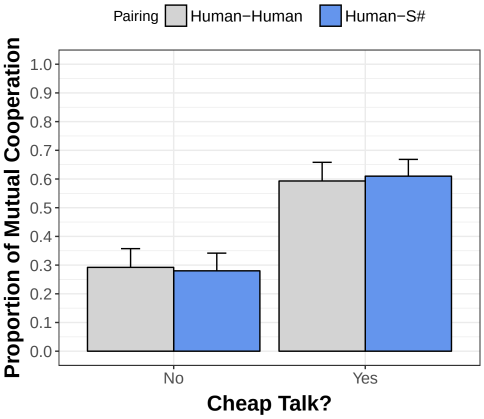

Figure 3: Proportion of mutual cooperation achieved by Human-Human and Human-Bot (where the bot was controlled by the algorithm S#) pairings across several repeated games as reported in prior work. Figure adapted from *Crandall et al.* [5]. Error bars show the standard error of the mean.

Repeated games with cheap talk [5, 19] offer an even richer model of long-term interactions. These games are similar to repeated games, except that they allow players to send a set of costless signals (known as cheap talk) to each other prior to acting in each round. As we demonstrate in the remainder of this section, the ability to communicated via speech acts allows a player to create narrative by signaling what it plans to do in the future (express their intent), proposing joint solutions for the players to following together (proposals of shared intent), and reflecting on the results of previous rounds. Thus, by carefully selecting and combining together speech acts, players can use cheap talk to compose narratives of the players' relationship and joint experience, both past and future. Such narratives, if constructed properly, can enhance and sustain long-term human-robot relationships.

5.2 The Impact of Cheap Talk

Prior work illustrates the power of narrative in long-term human-robot interactions to enhance human-robot interactions in repeated games. One study [5], whose results are summarized in Figure 3, found that two people typically do not form cooperative relationships in repeated games in which they cannot engage in cheap talk. However, when two people are allowed to send each other messages prior to acting in each round, the amount of mutual cooperation doubled. A nearly identical trend was observed, across the same scenarios, when a human was paired with a bot following

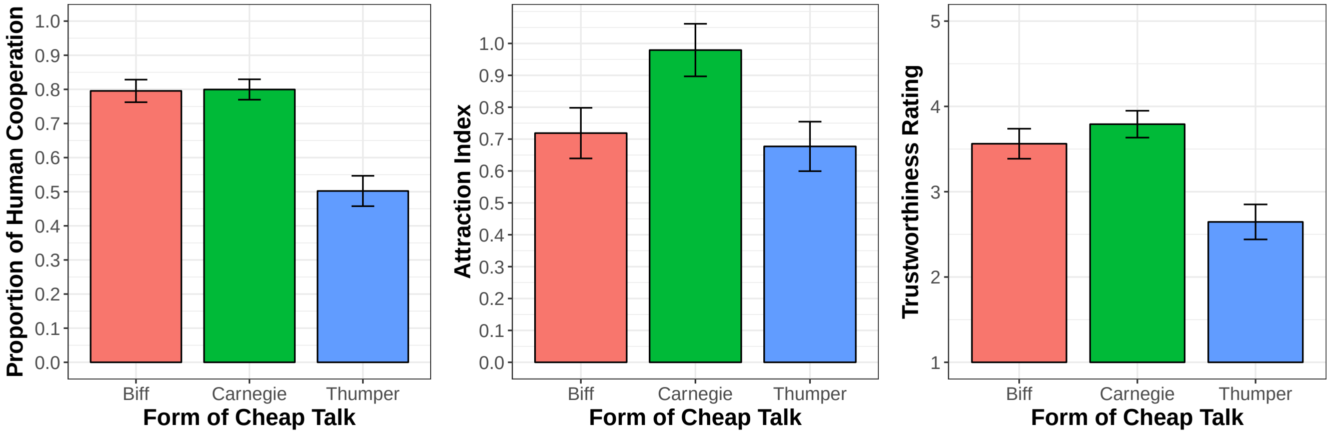

Figure 4: Selected results of a user study reported by *Oudah et al.* [19]. Results are shown only for the case in which the bot used the algorithm S# to select actions and speech acts. (a) The proportion of rounds that the human partner cooperated for each form of cheap talk. (b) The attraction index measuring people's affinity for interacting with a bot for each form of cheap talk. (c) The degree to which people thought there partner (the bot) was trustworthy for each form of cheap talk. In all plots, error bars show the standard error of the mean. See *Oudah et al.* [19] for details about how each variable was measured and calculated.

the algorithm S#, a result that demonstrates that bots likewise can and should use cheap talk to improve their long-term relationships with people.

In subsequent work, Oudah et al. [19] considered how using different kinds of speech acts (or form of cheap talk) impact not only a bot's ability to forge cooperative relationships with people, but also people's desires to continue to interact with the bot when given the choice. We focus on three different forms of cheap talk, each of which used a different form of cheap talk generated using the S# algorithm. These forms of cheap talk were Carnegie (which can be loosely defined as a friendly form of communicating), Biff (which can be loosely defined as an unfriendly way of communicating), and Thumper (which did not use any cheap talk).

Oudah et al. [19] found, via user study, that both Carnegie and Biff influenced people to cooperate with it better than did Thumper (Figure 4a). This caused Carnegie and Biff to receive higher rewards across the full repeated game than did Thumper. However, the cheap talk generated by Biff, as well as the lack of narrative produced by Thumper, caused people to be less disposed to want to interact with these bots further (as opposed to Carnegie; see Figure 4b). Post-experiment surveys revealed that both Biff and Thumper were not liked by study participants as much as Carnegie, and that Thumper was viewed as being significantly less trustworthy than both Carnegie and Biff (Figure 4c).

These examples from prior work illustrate that cheap talk can play a critical role in preserving and enhancing longitudinal human robot interactions. In particular, the bots were able to use cheap talk to (a) increase the amount of cooperation between humans and bots, (b) make humans more inclined to continue to interact with them, and (c) increase its perceived trustworthiness. In the next section, we consider the degree to which this cheap talk relates to the construction of narrative.

5.3 Cheap Talk as Narrative

Recall that our goal is to learn to produce an intentional narrative that avoids or recovers from undesirable equilibria or outcomes, such as misuse, disuse, and overuse traps. We now consider to what degree the cheap talk used in the previously discussed prior work does this. In particular, we focus on how and to what degree S#, the algorithm used to select speech acts in that prior work, produces the elements of narrative discussed by Bruner [3].

We begin by discussing how S# develops the landscape of action, which deals with episodes and the construction of the plot. We then discuss the landscape of consciousness, which deals with intentions, attribution, and repair.

5.3.1 Developing the Landscape of Action. In repeated games with cheap talk pair, the two players are the subject of the narrative, which is given voice by the speech acts selected by the bot. Each player has the high-arching goal to maximize its individual individual payoffs. These payoffs depend on the actions taken in each round of the game by the players. We can view each round as a separate episode of interaction, and the complete relationship is the sequence of these interactions.

The algorithm S# [5] creates an internal memory of the interaction using a finite-state machine. This finite-state machine encodes the algorithm's internal state combined with game-invariant events that are triggered by the actions taken by the players in the game. In a sense, this finite-state machine encodes the episodic memory of the agent (i.e., the memory of recent episodes), viewed in the context of S#'s current internal state, that have most recently occurred. S# also compiles into long-term memory a rudimentary summary of all past episodes by counting the number of times that each game-invariant event has occurred.

To voice a narrative of the interaction, S# uses its finite-state machine to identify the category of speech acts it should voice.

Table 2: Speech acts for the form of cheap talk called Carnegie**. See *Oudah et al.* [19].**

ID	Speech act
0	Let's always play <solution>.
1	Let's alternate between <solution> and <solution>.
2	This round, let's play <solution>.
3	if we can agree, we'll both benefit.
4	let's explore other options that may be better for us.
5	good idea. as expected from a generous person like u. I accept your proposal.
6	good proposal. if u show that u are trustworthy, I will consider accepting it in the future.
7	a fairer proposal would work to your benefit.
8	your payoffs can be higher than this.
9	what u did is totally understandable, though it will not benefit u in the long run.
10	in the next round comes the expected penalty, but we can then return to cooperating.
11	I'm really sorry I had to do that.
12	let's move on. I am sure we can get along.
13	excellent! Thanks for cooperating with me.

The specific speech acts selected for these categories based on both the form of cheap talk being used and the rudimentary summary of episodes that S# stores in its long-term memory (see Oudah et al. [19] for details). Example speech acts for Carnegie [19] are shown in Table 2.

The sequence of speech acts produced by S# using this mechanism voices the plot of the human-robot interaction. It does so in two ways. First, the speech acts generated by S#' in each round often reflect on previous (typically recent) episodes. For example, after a round in which S# receives a satisfactory payoff, S# voices speech act #13 ("*excellent! Thanks for cooperating with me.*"). Second, S#'s speech acts often foreshadow the plot of future episodes. As an example, after being exploited in a round, S# uses speech act #10. This speech act not only foreshadows what will happen in the next round, but also hints at potential outcome of subsequent rounds. Furthermore, after punishing its associates defection in the next round, S# voices speech act #11 ("*I'm really sorry I had to do that.*"). This sequence of speech acts links together causally related episodes to create a richer narrative of the interaction.

5.3.2 Developing the Landscape of Consciousness. Recall that Bruner's *landscape of consciousness* refers to the part of the narrative that answers questions of what the subjects of the narrative know, think, and feel. In particular, this aspect of the narrative reveals the intentions of the agents, their attributions of blame or credit when intentions are not met, and their attempts to repair damage when the intentions of one or more of the subjects of the narrative are not satisfied.

Intentions are an important part of narratives generated by S#. The individual intention of each player is to maximize it's own payoffs throughout the course of the repeated game. When using Carnegie as its form of cheap talk, S# acknowledges the other player's intention repeatedly (for example, see speech acts 3, 4, 7, 8, and 9). Biff, on the other hand, focus its narrative more on its own intentions [19].

Another way in which S# invokes intention in its narratives is in the form of shared intent. For example, S# sometimes expresses, via cheap talk, a desire to cooperate by specifying a cooperative solution (using speech act #0) or a desire to find a fair solution (speech act #7). We note that, in other results presented by Oudah et al. [19] but not reviewed here, the ability to express shared intent in this form appears to be a substantial part of the bot's ability to enhance and preserve relationships, as the absence of this capability produced substantially inferior results in repeated games.

When the intentions of the players are not satisfied, narrative often involves blame (or other forms of attribution). In repeated games, players can either put the blame on themselves, their partner, or the difficulty of the game (produced by the conflicts present in the payoff matrix). The way that S# ascribes blame depends on the form of cheap talk used. Carnegie typically expresses blame indirectly. For example, after its partner does not cooperate as expected, it indirectly points out its partner's error (using speech acts #9-10). On the other hand, Biff takes a more direct approach in the same situation by saying "*selfish traitor! you've treated me very unfairly.*" We anticipate that these alternative manners for communicating blame were a primary cause for differences of human opinion toward Biff and Carnegie found by *Oudah et al.* [19].

Finally, once blame has been voiced or inferred in the narrative, an important aspect of the narrative is to repair. S# does this using speech acts #11 and #12. More specifically, once S# forgives its partner for having violated a shared intention, it states, "*let's move on. I am sure we can get along.*" This speech act is designed to move the relationship on from the past in a new direction, and hence can be a powerful aspect of the narrative. While the results of the study reported by *Oudah et al.* appear to indicate that this repair is largely successful with respect to returning the players to a state of mutual cooperation, it seems possible from the results that the repair is not complete in the case of Biff (as indicated by low desire of people to want to continue to interact with Biff; Figure 4b).

6 CONCLUSION

Our position is that robots and robot services can use effective narrative to preserve and enhance long-term human-robot interactions. In this paper, we have established a metaphor equating long-term HRI as a dynamic trajectory through narrative space. In so doing, we identified and discussed the important elements of narrative. We then appealed to prior work in repeated games [5, 19] to illustrate that narrative produced by cheap talk can be utilized to overcome conflict and shortcomings in longitudinal interactions between a person and a bot. We believe that future work should better identify how to create narratives that allow robots to avoid undesirable equilibria, including the overuse, misuse, and disuse of robots, in human-robot relationships designed to endure over long periods of time.

REFERENCES

[1] AD Baddeley. 2010. Primer: Working memory. *Current biology* 20, 4 (2010), 136–140.

[2] Yochai Benkler. 2006. *The wealth of networks: How social production transforms markets and freedom.* Yale University Press.

[3] Jerome S. Bruner. 2009. *Actual minds, possible worlds.* Harvard University Press.

[4] Martin A Conway. 2009. Episodic memories. *Neuropsychologia* 47, 11 (2009), 2305–2313.

[5] J. W. Crandall, M. Oudah, Tennom, F. Ishowo-Oloko, S. Abdallah, J. F. Bonnefon, M. Cebrian, A. Shariff, M. A. Goodrich, and I. Rahwan. 2018. Cooperating with Machines. *Nature Communications* 9, 233 (2018).

[6] Nathan Crilly. 2011. The design stance in user-system interaction. *Design Issues* 27, 4 (2011), 16–29.

[7] Lisa Cron. 2012. *Wired for story: the writer's guide to using brain science to hook readers from the very first sentence.* Ten Speed Press.

[8] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3 (2003), 143–166.

[9] Michael A Goodrich and Alan C Schultz. 2007. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction* 1, 3 (2007), 203–275.

[10] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527.

[11] Curtis M Humphrey and Julie A Adams. 2009. Robotic tasks for chemical, biological, radiological, nuclear and explosive incident response. *Advanced robotics* 23, 9 (2009), 1217–1232.

[12] Troy Dale Kelley. 2014. Robotic dreams: A computational justification for the post-hoc processing of episodic memories. *International Journal of Machine Consciousness* 6, 02 (2014), 109–123.

[13] William Labov and Joshua Waletzky. 1967. Narrative analysis. Essays on the verbal and visual arts, ed. June Helm, 12-44. (1967).

[14] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[15] Carlos León. 2016. An architecture of narrative memory. *Biologically Inspired Cognitive Architectures* 16 (2016), 19–33.

[16] B. F. Malle. 2006. The Actor-Observer Asymmetry in Attribution: A (Surprising) Meta-Analysis. *Psychological Bulletin* 132, 6 (2006), 895–919.

[17] Bertram F Malle and Joshua Knobe. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology* 33, 2 (1997), 101–121.

[18] B. F. Malle, J. M. Knobe, and S. E. Nelson. 2007. Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology* 93, 4 (2007), 491âĂŞ514.

[19] M. Oudah, T. Rahwan, T. Crandall, and J. W. Crandall. 2018. How AI Wins Friends and Influences People in Repeated Games with Cheap Talk. In *Proceedings of the 32nd National Conference on Artificial Intelligence*.

[20] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[21] Francesca Polletta, Pang Ching Bobby Chen, Beth Gharrity Gardner, and Alice Motes. 2011. The sociology of storytelling. *Annual Review of Sociology* 37 (2011), 109–130.

[22] Nicolas Szilas. 2015. Towards Narrative-Based Knowledge Representation in Cognitive Systems. In *OASIcs-OpenAccess Series in Informatics*, Vol. 45. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.